

Brokering Knowledge from Laboratory Experiments in Evidence-Based Education: The Case  
of Interleaving.

Paul Rowlandson and Adrian Simpson\*

School of Education

Durham University

\*Corresponding author: Adrian Simpson, School of Education, Durham University, Stockton  
Road, Durham, DH1 3LE, United Kingdom. Email: [adrian.simpson@durham.ac.uk](mailto:adrian.simpson@durham.ac.uk)

Accepted May 2024

To appear in *British Education Research Journal*

Brokering Knowledge from Laboratory Experiments in Evidence-Based Education: The Case  
of Interleaving.

### Abstract

The turn to 'evidence-based education' in the past three decades favours one type of evidence: experiment. Knowledge brokers ground recommendations for classroom practice on reports of experimental research. This paper distinguishes *field* and *laboratory* experiments, on the basis of control and precision of causal ascription. Briefly noting problems with knowledge brokers' extrapolating from field experiments, the paper's main focus is on extrapolating from laboratory experiments, using the case of 'interleaving'. It argues knowledge brokers often extrapolate from laboratory experiments as if they are field experiments. By considering both laboratory and 'extra-lab' interleaving studies, it suggests an alternative extrapolation – creating laboratory effects in the classroom – has little pedagogical value. The conclusion suggests focussing on mechanisms, contexts and outcomes as a more useful basis for brokering pedagogical knowledge from laboratory experiments.

#### Key Insights:

*What is the main issue that the paper addresses?*

The paper focusses on how experimental research knowledge is brokered for practice. It highlights substantial problems with the way knowledge brokers transport phenomena from field experiments. Then, using the case the 'interleaving' in category learning research, it explores how knowledge brokers transport phenomena from cognitive science to the classroom.

*What are the main insights that the paper provides?*

Transporting knowledge from field and laboratory experiments to classroom practice cannot be achieved directly. While in scientific disciplines, laboratory phenomena can be transported through engineering, this route is not applicable for pedagogy. Focussing on the mechanisms, contexts and outcomes of experimental research is more promising for brokering knowledge for practice.

## **Brokering Knowledge from Laboratory Experiments in Evidence-Based Education:**

### **The Case of Interleaving.**

Social policy and professional practice, including education, are increasingly subject to demands that they should be informed by research. In many contexts, this has been translated into the rhetoric of 'evidence-based policy'. This underpins education policy in many countries, including the US's 'No child left behind' (2002) and 'Every student succeeds' (2015) acts and the formation of the Institute for Education Sciences (IES); Australia's 'Productivity Commission' (2016) and Ireland's 'Programme for Government' (Department of the Taoiseach, 2020).

In the United Kingdom, 'evidence-based education' (EBE) might be traced to an influential speech to the Teacher Training Agency which argued education should look to the apparent success of medicine for inspiration; becoming a 'research based profession' (Hargreaves, 1996). The UK's EBE movement has been integrated into policy, with government designating the Education Endowment Foundation (EEF) as a 'what works clearinghouse' aimed at synthesising existing evidence, promoting evidence use and commissioning studies to generate new evidence (Edoald & Nevill, 2021).

While some suggest the roots of EBE predate evidence-based medicine (Baron, 2018), it is often improved medical outcomes which are used to justify similar approaches in education (Slavin, 2002).

Critical to these approaches is a particular interpretation of 'evidence'. While definitions vary, there is a common element: experiment. Davies (1999) claimed:

For those who ask questions such as 'does educational method (or health care intervention) x have a better outcome than educational method (or health care

intervention)  $y$  in terms of achieving outcome  $z'$ , evidence consists of the results of randomised controlled trials or other experimental and quasi-experimental studies.

(p. 114)

and Slavin (2020) argued:

Evidence of effectiveness is defined as evidence from rigorous experiments in which students experiencing experimental programs are compared over significant periods (say, a semester or more) to those using traditional control methods in terms of gains on valid measures of achievement or other outcomes. Ideally, students, teachers, and/or schools are assigned at random to experimental or control treatments. (p.

22)

For example, EEF-funded research overwhelmingly takes the form of randomised controlled trials. Its impact is highly influential: this single, UK-focussed charity reportedly commissioned nearly 20% of education trials worldwide in the past 10 years (Edoald & Nevill, 2021, p. 49). That influence arguably monopolises the policy landscape: “The EEF are so successful that they are now effectively functioning as a gatekeeper, deciding whose knowledge counts” (Innes, 2023, p. 13).

However, published reports of experiments may not influence practitioners directly (Dagenais et al., 2012). While myths persist of isolated teachers, determining practice alone (Lortie, 2020), ideas for practice usually derive from interactions with colleagues, through continuing professional development, reading professional journals, books and, increasingly, online sources and social media (Torphy et al., 2020). Critically, teachers associate credibility and trustworthiness with how this material is grounded in academic research (Gleeson et al., 2022). That is, professional literature authors are seen as

‘knowledge brokers’, distilling research and translating it into accessible language (Rycroft-Smith, 2022).

The focus of this paper is on the mechanisms through which experimental research becomes promoted as knowledge for practice. The paper is not intended as a systematic review of interleaving research or of knowledge brokering. Instead, its contribution is as a theoretical discussion of how laboratory studies come to be used as evidence for policy, illustrated through an examination of one particular strand of experimental research.

Two forms of experiment are distinguished: field and laboratory studies. The first of these, the field experiment, is discussed briefly. In this case, knowledge is often brokered through direct extrapolation and the paper notes concerns with the underpinning assumption that the causal roles identified in field experiments can transport to classroom practice in this direct manner,

The main focus, however, is on how laboratory experiments are brokered for practitioners. This is illustrated with the case of the ‘interleaving effect’. This effect is the apparent improvement in classification when different categories’ exemplars are encountered sequentially (interleaved, e.g. ABCABCABC...) rather than together (blocked, e.g. AAABBBCCC...). Knowledge brokers’ claims for interleaving are discussed, before the focus shifts to the underpinning laboratory studies which have provided a robust and well replicated evidence base for an interleaving effect and for the circumstances in which that effect is generated, suppressed or reversed.

Laboratory experiments are characterised by careful control, allowing researchers to ascribe cause more precisely. The paper discusses two ‘extra-lab’ experiments in which there is a “controlled relaxation of control” (in the sense of Nagatsu & Favereau, 2020).

When some of the artificiality of the laboratory is exchanged for realistic classroom features, the effect failed to appear.

The paper claims these laboratory and extra-lab experiments highlight a fundamental problem with brokering knowledge from such studies. Knowledge brokers often treat laboratory experiments as field experiments: extrapolating directly. Such direct extrapolation is problematic, even for field experiments, but the interleaving example suggests directly brokering knowledge from laboratory experiments is particularly unjustified. It concludes that knowledge brokers need to treat evidence from laboratory experiments differently. While it may be possible to engineer laboratory effects like interleaving in classrooms, more successful practice may come from understanding underlying mechanisms and contexts which are teased out by laboratory experiments.

### **Experiments in Evidence-based Policy**

In discussing evidence-based economics, Nagatsu and Favereau (2020) distinguished two strands of experiment – both involving random allocation to treatments – characterised by levels of control. The field experiment grew from concerns about evaluating policy in realistic contexts, while the laboratory experiment grew from psychological traditions of assessing individual decision making. In both cases, random allocation uses statistical methods to ascribe a causal role on differences in outcomes to post-allocation differences in treatment. Laboratory experiments permit more careful control over the nature of that difference in treatments, treatment adherence and measurement.

While the paper focusses on brokering knowledge from laboratory experiments, it is useful to contrast this with field experiments.



The EEF and IES support resource-intensive, large-scale evaluations of educational programmes, aimed at identifying ‘what works’. For example, the EEF funded an evaluation of a ten-week programme of highly scripted arithmetic lessons (Nunes et al., 2018). These were delivered by specialist, trained teaching assistants to small groups of primary school children struggling with mathematics. Groups were randomly assigned to the programme or to continue with normal teaching, with performance on a quantitative reasoning test as the primary outcome. The mean intervention group score was higher than the comparison, and the study was subsequently promoted to teachers as evidence of the value of high-quality teaching assistant support in primary school mathematics (Hodgen et al., 2020).

That is, for field experiments brokering knowledge may be relatively direct. The relative success of the group with trained teaching assistants was taken as evidence for recommending future interventions with this feature: it ‘worked there’, so will ‘work here’. In some cases, particular field experiments are taken as direct grounds for knowledge brokers’ recommendations; in others ‘meta-analysis’ and ‘meta-synthesis’ combine results from multiple experiments to rank order general forms of practice as ‘good bets’ for improving learning outcomes (Higgins et al., 2022).

Brokering knowledge is an issue of external validity - the extent to which “the causal relationship holds over variation in persons, settings, treatment, and measurement variables” (Shadish, et al., 2002, p. 20). Among other concerns, extrapolating from a field experiment to a given classroom involves addressing the ‘black box’ nature of the causal relationship. The experiment might rigorously establish that the total of all post-allocation differences played a causal role on the average difference in outcomes. However, it does not identify which combination of post-allocation differences came together to create the effect; whether some acted to decrease the effect; which participants would have been

positively or negatively impacted by being allocated to the other treatment nor what contextual factors present may have facilitated or inhibited effects (Cartwright & Hardie, 2012).

Nunes et al.'s (2018) arithmetic field experiment combined many different elements including scripting, small group learning, additional teaching time, tutor training, use of multiple representations, participants with particular support needs but good group working skills, a weakly specified 'business as usual' control group etc. As knowledge brokers, Hodgen et al. (2020) took the study as evidence for a general positive causal role for high-quality teaching assistant support. Yet there are no direct grounds from the experiment for identifying which combination of features contributed to the overall average difference in outcomes: it is possible that teaching assistants played a negative role which was outweighed by the positive roles of additional teaching time and small group learning.

Warrants for brokering knowledge for practice directly from field experiments must, then, rely on 'high fidelity': "Even granting projectability [the extent to which past instances can be taken as guides for future ones], unless one replicates the whole set of post-allocation differences with exact fidelity, there is no evidence that similar effects will occur" (Joyce & Cartwright, 2020, p. 1070).

So, brokering knowledge for practice from field experiments is often taken to be direct, but is beset with difficulties about transporting causes. The bulk of this paper is focussed on brokering knowledge from another form of EBE: laboratory experiments. The careful control of a laboratory experiment reduces the 'black box' problem, allowing researchers to identify causes more precisely and, across a sequence of experiments, tease out circumstances leading to the generation, suppression or reversal of effects.

### Laboratory Experiments in EBE

As well as field experiments, knowledge brokers draw on laboratory experiments to recommend particular classroom practices. The Deans for Impact (2015) report makes recommendations for teachers, often citing cognitive science laboratory experiments. For example, encouraging students to “identify and label the substeps required for solving a [multi-step] problem” (Deans for Impact, 2015, p4) is grounded on two papers. These report six laboratory experiments, predominantly involving psychology undergraduate students with little prior relevant knowledge, solving multi-step Poisson distribution problems, with labelling of steps being carefully controlled (Catrambone, 1996; 1998).

Clearly, the original research aim was not to provide direct pedagogical advice. Experiments aim “to create, produce, refine and stabilize phenomena” (Hacking, 1983, p. 230) using researchers’ clever arrangements of apparatus, material and measuring instruments. The work of experimenters such as Catrambone is to carefully control features between experiments such as the nature and number of steps in a problem, how it is labelled etc., to create the phenomenon, turn it off and even reverse it.

In doing so, one might argue that science discovers laws which are generally applicable elsewhere. Hacking (1983) argued against this, even in pure sciences like physics. While experimenters and technicians bring together equipment and material to generate a phenomenon in ever purer form, that phenomenon might not be immediately available elsewhere. In discussing the ‘Hall effect’ in Physics, Hacking notes:

I suggest ... that the Hall effect does not exist outside of certain kinds of apparatus.

Its modern equivalent has become technology, reliable and routinely produced. The effect, at least in a pure state, can only be embodied by such devices.

That sounds paradoxical. Does not a current passing through a conductor, at right angles to a magnetic field, produce a potential, anywhere in nature? Yes and no. If anywhere in nature there is such an arrangement, with no intervening causes, then the Hall effect occurs. But nowhere outside the laboratory is there such a pure arrangement. (Hacking, 1983, p. 226)

Hacking's contention is that transporting a result from the laboratory involves creating lab-like conditions on the world. That is, once scientists have the level of control required to refine the phenomenon, engineers can create technology with those conditions to exploit it. For example, a Hall effect sensor, with just the right configuration of components, can exploit the effect to create a speedometer for a car. That requirement to impose strongly on the world to create conditions to exploit an effect ('lab-ifying' a piece of the world) likely applies to many scientific phenomena.

While the conditions under which an effect can be generated is one form of useful knowledge from a sequence of laboratory experiments, scientists can also generate theory: an understanding about the mechanisms which resulted in the observed effect and which might be available to be exploited elsewhere.

Knowledge brokers thus have two valid routes for extrapolating evidence from the laboratory. First, detailing conditions under which the phenomenon can be generated so engineers can intervene to 'lab-ify' some part of the world to exploit it. Second, explaining the mechanisms – and contexts in which those mechanisms might work – so practitioners can look at their own contexts to see if the mechanisms might be exploited to achieve a desirable outcome.

This paper argues that knowledge brokers in EBE often take neither of these routes. Instead, they treat knowledge from laboratory experiments in the same way as knowledge

from field experiments; trying to extrapolate directly from experimental treatments to classrooms. The paper further argues that 'lab-ifying' the world, while it might generate the effect in a classroom setting, has little pedagogical value. Thus, only a focus on mechanisms and contexts may have value for brokering knowledge from laboratory experiments to classrooms.

This will be explored through the case of interleaving: mixing the order in which examples from different categories are encountered, in contrast to blocking in which examples from the same category are encountered together. First, recommendations for practice from knowledge brokers are illustrated and a particular issue about the notion of a category is highlighted. The underpinning reports from laboratory experiments are discussed, noting the characteristics of careful control. Two recent experiments are discussed in which a small amount of that careful control is exchanged for increased authenticity (so called 'extra-lab' experiments). Taken together, the laboratory and extra-lab experiments highlight two difficulties with the knowledge broker literature on interleaving: first, it conflates two separate sets of experiments; second, it extrapolates causal roles via the wrong route. Finally, the paper explores what knowledge brokers might obtain by focussing on mechanisms and contexts.

### **Knowledge Brokers and Interleaving**

Professional education literature is a knowledge brokering system for teachers. It often recommends interleaving as a strategy, referencing laboratory experiments. Deans for Impact (2015) argued "if students are learning four mathematical operations, it's more effective to interleave practice of different problem types, rather than practice just one type of problem, then another type of problem, and so on" (p.2). Barton (2018) noted "The Interleaving Effect contrasts a 'blocking' approach, whereby students study the same

type of material over and over again before moving on to a different type of material, against an 'interleaving' approach, where students practise all of the problems in an order that is more random and less predictable. The latter approach has been found to enhance learning and transfer." (p.410)

Some literature focusses on how interleaving supports learners distinguishing between categories. For example, Weinstein et al. (2018) claimed "Interleaving allows the learner to better distinguish between different concepts" (p. 96) and Brown et al. (2014) suggested "When you're adept at extracting the underlying principle or rules that differentiate different types of problems, you're more successful at picking the right solutions in unfamiliar situations. This skill is better acquired through interleaved and varied practice than massed practice" (p. 4). In promoting interleaving, Barton (2018) argued "presenting related concepts together forces students to distinguish between them, and hence benefit from interleaving and the power of non-examples" (p. 417). Agarwal and Bain (2019) suggested, "In order to encourage discrimination, the key is mixing up similar ideas" (p. 112)

Some knowledge brokers maintain that interleaving supports the building of connections: "interleaving helps students to make connections between different topics or categories" (Weinstein et al., 2018, p. 84); "as well as being able to spot the differences between each topic, interleaving also helps students to focus on the similarities that they previously might have not been aware of" (InnerDrive, n.d.).

Many knowledge brokers suggest interleaving has long term effects on performance. For example, Brown et al. (2014) claimed "research shows unequivocally that mastery and long-term retention are much better if you interleave practice than if you mass it" (p. 50). Also, Busch and Watson (2019) suggested "a growing body of evidence ... has found that

interleaving types of problems within a subject helps improve long-term retention, recall and performance" (p. 36).

It might appear as if the distinction between interleaved and blocked sequencing is unproblematic. The latter involves items of the same category being presented sequentially and the former mixes items from different categories. However, what counts as interleaving or blocking depends on what makes categories coherent: i.e., what makes otherwise distinguishable objects able to be treated equivalently. Markman (1989) argued this usually involves a highly complex process of constrained induction.

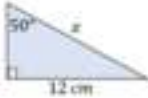
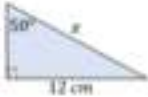



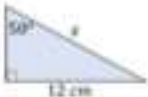
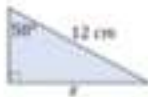
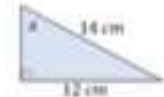
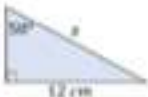


It is not always immediately clear whether a sequence of items belongs to one category (i.e., blocked) or to multiple categories (i.e., interleaved). In figure 1, sets of items are, on one reading, equivalent and, on another, not.

That is, when trying to inductively construct categories from exemplars, what counts as interleaved or blocked can be ill-defined. Moreover, the extension of inductively learned category can be unclear – if all exemplars of 'quadrilateral' are convex, simple and planar, a learner may subsequently exclude concave, intersecting or non-planar items.

Knowledge brokering literature is not clear about the nature of the categories which would benefit from interleaving: would sequences mixing mathematics, geography and French benefit from interleaving as much as sequences mixing sine, cosine and tangent?

**Figure 1**

*Illustration of exemplars as members of a single or of multiple categories.*

	Exemplars			Single category	Multiple categories
a		What is the French word for "mollard"?	Provide a series of arguments for and against the use of the death penalty as a way to punish criminals.	Homework	Subject discipline
b		Calculate the mean of these numbers: .8, 7, 9, 4, 7, 6	Work out: $\frac{3}{5} + \frac{4}{5}$	Mathematical	Topic area
c	  			Trigonometrical	Trigonometrical ratio
d	  			Sine ratio solution	Solution process
e	  			Solution process	Orientation

Given the professional literature is unclear about what makes sequences interleaved or blocked, and therefore what makes them candidates for the interleaving effect, it may be useful to review the underpinning research literature to examine the nature of items in these experiments and other ways in which control is exercised in laboratory settings.

**Laboratory Experiments in Interleaving.**

There is a large, robust and well replicated cognitive science literature exploring the interleaving effect. The seminal laboratory study presented participants with landscapes and skyscapes painted by twelve relatively unfamiliar artists, each labelled with the artist's name (Kornell & Bjork, 2008). Some artists' works were blocked within the sequence,



others interleaved. Participants were tested on their ability to select the correct artists on subsequent unfamiliar paintings. Nearly 80% of participants were better at classifying for artists whose work had been interleaved. Despite this, over 70% of participants claimed they learned more from blocked presentation.

Consistent with Nagatsu and Favereau's (2020) idea of careful control in laboratory experiments, studies following Kornell and Bjork (2008) took advantage of the artificial, relatively noiseless laboratory environment; controlling features to more precisely identify circumstances where interleaving resulted in better classification than blocking. For example, Zulkipli and Burt (2013) controlled both sequencing of paintings and time between items, finding that while interleaving led to more accurate classification than blocking, there was no interaction with inter-item timing. This suggested the classification advantage did not result from temporal spacing of category exemplars which was a possibility in Kornell and Bjork's study: inevitably exemplars from a given category (such as 'A') are more spaced out in time in interleaved presentations (e.g. 'ABCABCABC...') than blocked (e.g. 'AAABBBCCC').

In their second experiment, Zulkipli and Burt (2013) controlled sequencing and between-category similarity – using highly artificial images instead of paintings – finding that when categories were easy to distinguish, blocking led to better classification and when categories were hard to distinguish, interleaving was better.

This approach of controlling experimental features to identify how sequencing affects categorisation has resulted in a rich literature encompassing photographs of birds (Birnbaum et al., 2013), abstract images (Eglington & Kang, 2017), text (Sana, Yan & Kim, 2017), sounds (Zulkipli, 2013) etc.

Nevertheless, a second, distinct group of studies is often included in discussions of interleaving effects. Rather than associating category names with exemplars, participants practise different types of procedures. For example, Rohrer and Taylor (2007) reported on students practising unfamiliar volume formulae for four different solids such as a spheroid and spherical cone. Practice was either blocked (e.g. all spheroid problems together) or mixed (alternating different solids). Subsequent performance was lower on average for mixed practice.

Similarly, Rohrer et al. (2014) gave school children practice problems from four categories (slopes, linear graphs, proportions and linear equations) across ten assignments in a nine-week period. Again, those who encountered practice questions in blocks had poorer average post-test results.

Despite the superficial similarity of category learning studies (such as Kornell & Bjork, 2008), and repeated practice studies (such as Rohrer et al., 2014), they are critically different. The former, involve recognising items' category membership and recalling category names. These experiments tease out the important role played by between-category similarity. The latter, repeated practice studies, may involve recognition, but a mathematical procedure must be retrieved and followed once the problem type is recognised. Often, category similarity plays no role: in Rohrer et al. (2014) the mathematical categories were so distinct, the researchers found participants never confused them.

The effect in repeated practice experiments appears more plausibly explained by problems being spaced-instead-of-massed, rather than being interleaved-instead-of-blocked (Foster et al. 2019). Time between practice may prompt retrieval of formulae from memory; each successful retrieval strengthening that memory. Moreover, the intellectual

demands of massed practice – when same procedure is repeated – may be low, leading to poor attention; while practice spaced out by anything (not just another problem of a similar type) may maintain attention. The characteristic of interleaving – placing items of one category against another – does not appear to be a critical element of repeated practice experiments.

Interleaving studies involve multiple, potentially confusable categories. Since the seminal work of Kornell and Bjork (2008), a broad range of research has developed, reporting laboratory experiments of the relative efficacy of interleaved and blocked presentations on inductive category learning. Many studies have replicated the effect using very similar conditions to Kornell and Bjork (2008), some including the same set of paintings (e.g. Kang & Pashler, 2012; Metcalfe & Xu, 2016; Zulkipli & Burt, 2013). In the search for the circumstances where the effect is facilitated, suppressed or reversed, a variety of other materials and experimental designs has been used.

Nonetheless, these studies share characteristics: a passive training phase where labelled exemplars from different categories are presented in a particular sequence within a short period (often a few minutes). This is followed rapidly by testing where items are presented and participants select a category name from a list. Generally, the material has no educational relevance to the participants, who are often university psychology students engaging for course credit (Firth, Rivers & Boyle, 2021).

These features demonstrate the careful control characteristic of the laboratory experiment highlighted by Nagatsu and Favereau (2020). They reduce noise and enhance the ability to identify causal features. For example, passive presentation allows the researcher to control material and sequencing to see how these impact on the size and

direction of any effect. Short experiments, with training and testing phase together, reduces noise from participant fatigue and attrition.

The laboratory context also addresses the issue of what constitutes a category (as illustrated in figure 1). In the experiments, categories are determined by the task: categories are disjoint and the correct name is presented with each exemplar during the training phase. The testing phase is normally multiple choice, so classification can be facilitated both positively (belonging to a category) or negatively (not belonging to other categories).

The sequence of carefully controlled studies since Kornell and Bjork (2008) has allowed researchers to generate and test hypotheses about mechanisms accounting for the effects. Retrieval mechanisms – in which forgetting and ‘reloading’ strengthens memory (Bjork & Bjork, 2011) – appear to be more plausible accounts for spacing/massing effects in repeated practice studies (such as Rohrer et al., 2014). Temporal spacing of items, particularly if separated by another task involving attention, results in forgetting; so subsequent items require reloading. This mechanism may work for any separating task – it need not require the space between target category items to involve a second, potentially confusable category.

Unlike repeated practice studies, category learning experiments seem to involve different categories where learners are aggregating exemplars; abstracting relevant and irrelevant information across presentations; and checking deductions against new instances (Vlach & Sandhofer, 2013). Three mechanisms are proposed to account for interleaving effects: attention attenuation, discriminant contrast, commonality abstraction.

The first involves inattention being increased by familiarity. Consecutive presentations from the same category results in later occurrences receiving less processing

and being less well encoded (Gerbier & Toppino, 2015; Vlach & Sandhofer, 2013). Metcalfe and Xu (2016) found higher reported levels of 'mind wandering' for blocked items.

The discriminant contrast mechanism posits interleaving leads to more cross-category comparison opportunities and thus draws attention to discriminating features (Eglington & Kang, 2017).

However, as well as identifying distinguishing inter-category features, classification can be facilitated by identifying features common to a category. This is the proposed commonality abstraction mechanism: juxtaposing examples from the same category increases common feature salience (Sana et al. 2017).

These last two mechanisms appear in tension. Juxtaposing exemplars from the same category facilitates recognition of features, positively identifying membership. Juxtaposing exemplars from different categories facilitates discrimination. In different circumstances, these competing mechanisms will be more or less effective for classifying. This may explain moderating effects of inter- and intra-category similarity. Carvalho and Goldstone's (2014) experiments manipulated features shared between categories, or shared between exemplars within categories. Blocking outperformed interleaving with low within-category-similarity materials, with interleaving dominating for high between-category-similarity materials.

This illustrates how control available in laboratory experiments allows cause to be ascribed to more precisely defined differences in treatments. Subtle changes to experimental features, across a sequence of experiments, allows researchers to identify when the phenomenon can be generated, suppressed or reversed. This enables them to posit and test theories about mechanisms which create effects under different conditions.

For a knowledge broker, the question is how to extrapolate from those experiments to useful recommendations for practice. One approach to bridging the gap between research and practice is exchanging some of the careful control for features closer to intended practice – which Nagatsu and Favereau (2020) described as “controlled relaxation of control”.

### **Extra-lab Experiments – Controlled Relaxation of Control.**

The majority of interleaving laboratory experiments uses material of no relevance to the participants: often university psychology students, knowingly taking part in an experiment in an artificial environment. The control provided by restricting experiments on the three dimensions of relevance, participant and location can be exchanged for more apparently authentic values; relevant materials, school students and a school setting.

Rowlandson and Simpson (2023) reported on two extra-lab interleaving experiments in the context of learning mathematics. Many laboratory experimental features were retained: random allocation to treatment, short training and test phases with little delay between them. However, participants were secondary school students, the topic was educationally relevant (angle relations in parallel lines) and the setting was a school computer suite.

As well as blocking and interleaving, the experiments had a third treatment. Noting that simple induction of categories from exemplars is an uncommon pedagogical strategy, participants in a third group (‘exposition’) were told the defining features of each category. For example, alongside one image and category name, exposition group participants were told co-interior angles are on the same side of the transversal, inside the pair of parallel lines. Blocked group participants were shown a set of co-interior examples together, with

other angle relation categories blocked in a similar manner. The final group were shown the exemplars for different categories interleaved (see figure 2).

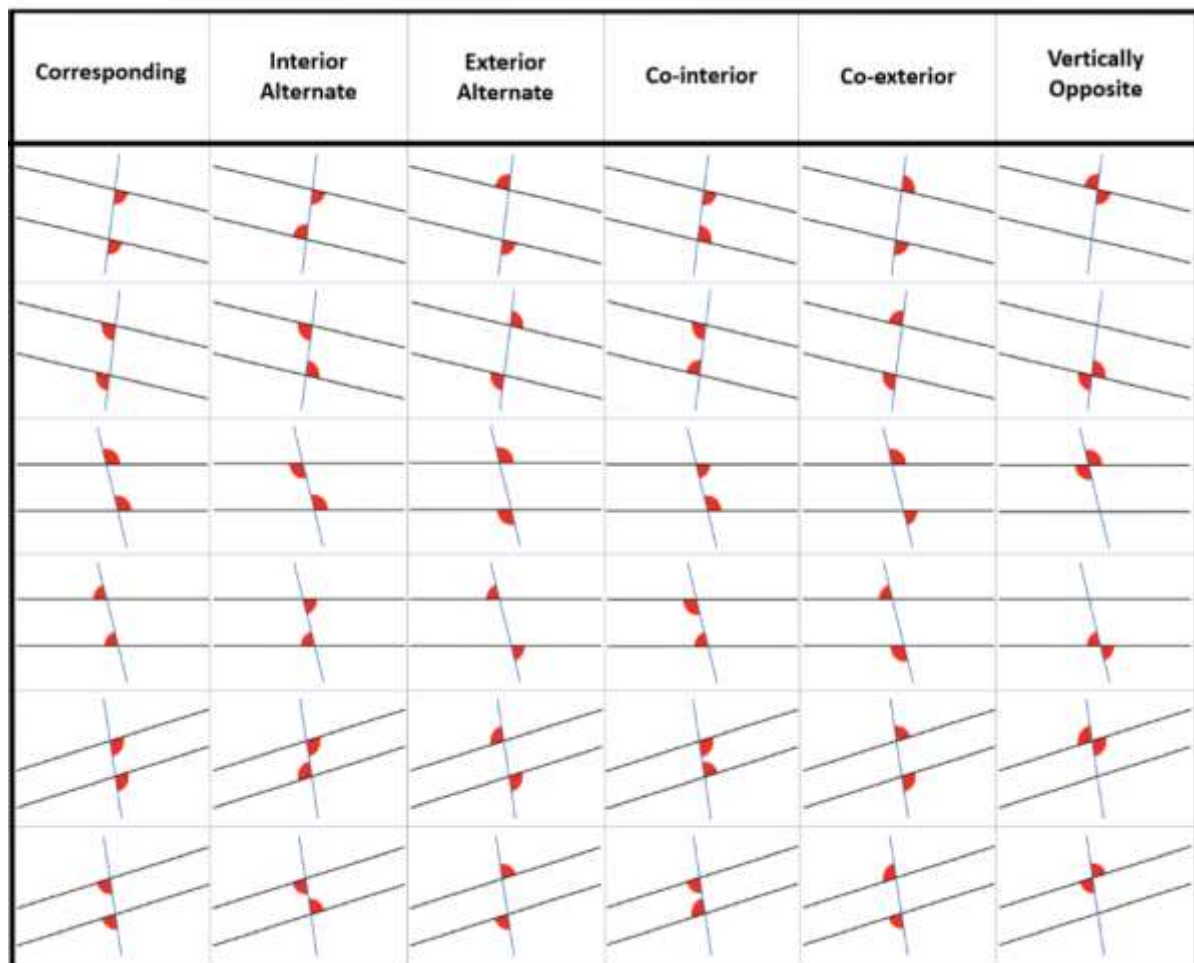
Testing involved classifying similar images. Despite being well powered for a study of this type, no effect was detected. Exploratory analysis highlighted some issues.

Classification of two categories (corresponding and vertically opposite) was high, while classification of the other four was around chance level. There was clear evidence that concentration levels decreased across the experiment and some participants were observed to be completely disengaged.

A second experiment addressed these issues. A larger, older and somewhat higher attaining group was recruited. This increased the power to detect smaller effects and addressed concerns about task difficulty and motivation. The training and test phases were shortened to maintain attention. The number of categories was reduced to four, removing the most easily distinguished (vertically opposite and corresponding) leaving those with the highest between-category similarity; considered most likely to support an interleaving effect.

**Figure 2**

*The training phase stimuli used for the blocked and interleaved groups in Rowlandson and Simpson (2023).*



Again, no difference was detected in the average score on the classification test.

However, in contrast to the first experiment, all three groups classified very accurately.

One suggestion from Rowlandson and Simpson (2023) is that the categories used were classical rather than natural (in the sense of Rosch, 1973). Angle categories are defined by a small number of easily identified features, while painting-artist categories are not. If a participant identifies defining features for angle images, classification is near perfect; if they do not, classification is close to chance levels. Category judgement mechanisms may be different for classical categories: rather than seeking holistic similarities/differences, participants may be seeking definitional features. In the extra-lab



experimental context, blocking, interleaving and exposition may have supported definitional feature identification equally effectively.

Whatever the reason, the controlled relaxation of control resulted in suppression of the interleaving effect in both experiments. Something about that relaxed control resulted in the effect disappearing compared to the laboratory experiments: the categories were no longer quite the right kind of categories, the participants were no longer quite the right kind of participants, the judgement mechanisms evoked were no longer quite the right judgement mechanisms etc.

Had Rowlandson and Simpson continued their experiments, tweaking materials, participants and tasks, they might have engineered just the right set of circumstances to create the interleaving effect (and perhaps an exposition effect too) in an extra-lab experiment. But to what end? They would have demonstrated the effect can be extrapolated at least a little way from carefully controlled laboratories, but that says little for pedagogy.

### **Three Routes to Extrapolation**

There are three routes to extrapolating from experiments to classrooms. In field experiments, knowledge brokers appear to extrapolate the treatment directly (it ‘worked there’ so it ‘will work here’). For laboratory experiments, if scientists and engineers identify the laboratory conditions in which the effect is generated, they can ‘lab-ify’: create lab-like conditions to generate the effect. Finally, identifying what experimental reports say about the mechanisms and their facilitating contexts might enable practitioners encountering those contexts to exploit those mechanisms.

The argument here is that knowledge brokers appear to take the direct route for laboratory interleaving experiments, which does not result in effective extrapolation. While

the second route – ‘lab-ifying’ – can result in effective extrapolation, it may not be suitable for education. The third route, then, may be the most effective for using knowledge from laboratory research to inform teaching.

### **Direct extrapolation**

Knowledge brokers often extrapolate from field experiments directly: features of the more successful arm of the evaluation are taken to be generally effective. For example, the positive evaluation of small group, scripted, teaching-assistant led arithmetic lessons (Nunes et al. 2018) is taken as direct evidence to recommend high-quality teaching assistant support (Hodgen et al. 2020).

In the case of interleaving, knowledge brokers often take a similar approach: citing laboratory studies where interleaving is more effective to directly recommend interleaving for practice.

The problems with direct extrapolation from field experiments outlined in Joyce and Cartwright (2020) – projectability, the role of agency, representativeness etc. – apply equally to direct extrapolation from laboratory experiments. However, there are additional problems for laboratory experiments, such as interleaving research.

Much professional literature conflates repeated practice and category learning research. Foster et al. (2019) showed time between items – rather than mixed sequencing of related category exemplars – facilitates improved procedure retrieval. Where retrieval practice studies use problems from different topics, there is no evidence of category confusion (Rohrer et al., 2014). Yet professional literature cites these studies interchangeably and suggests interleaving applies to practice problems (e.g. Barton, 2018; Brown et al., 2014; Weinstein et al., 2018). Taylor and Rohrer (2010) suggested this occurs

only when practice problems are so similar that students might confuse categories, but the practice retrieval effect appears independent of the interleaving effect.

In addition to conflating separate effects, professional literature fails to recognise that experiments do not always favour interleaving. While meta-analyses show weighted average effect sizes from studies favour interleaving over blocking, there is considerable heterogeneity (Brunmair & Richter, 2019; Firth et al. 2021). By manipulating experimental conditions, researchers can facilitate, suppress or reverse the effect. The claim “meta-analysis revealed a moderate overall interleaving effect (Hedges’  $g=0.42$ )” (Brunmair & Richter, 2019, p. 1029) says only that, so far, more (and clearer) studies have been conducted using one set of experimental conditions than using others. Nonetheless, professional literature tends to focus only on the advantage of interleaving, ignoring conditions which might favour blocking.

The Deans for Impact (2015) recommendation – “if students are learning four mathematical operations, it’s more effective to interleave practice of different problem types, rather than practice just one type of problem” (p.2) – combines both concerns. Interleaving may help distinguish categories of problems only for students who confuse, say, addition and division problems; on the other hand, memory for procedures may be facilitated by spacing rather than interleaving.

As well as concerns about conflating effects and overlooking critical support conditions, professional literature often omits any focus on the comparison condition in the research. The laboratory research does not show ‘interleaving works’. Instead, it suggests, in the right circumstances, interleaving works *better than blocking*.

Blocking in the research is an unusually pure version: exemplars are interchangeable elements of the category, and ordering within blocks is considered irrelevant. As noted

above, the terms ‘interleaving’ or ‘blocking’ depend critically on how instances are seen as the same or different. On one reading “tamiser, chacal, neveux, tandis” are blocked (French vocabulary); on another they are interleaved (different pronunciation rules) (Carpenter & Mueller, 2013). In whichever reading one has, however, “chacal, neveux, tamiser, tandis”, “chacal, tamiser, tandis, neveux” etc. are equally good examples of blocked/interleaved sequences.

This pure, interchangeable blocking is uncommon in classrooms. To support their contention that blocking is widespread educational practice, Rohrer, Dedrick and Hartwig (2020) presented the tasks “Simplify the expressions:  $4x + 3 - 9x$ ;  $5 + 3.2n - 6 - 4.8n$ ;  $2y - 5(y - 3)$ ;  $\frac{1}{2}(8b + 3) + 3b$ ” (p.876) as blocking. However, the question setter likely did not see these expressions as interchangeable elements of a common category, but as increasing in difficulty: i.e., neither blocked nor interleaved.

Teachers may draw on a wide range of principles to sequence examples, such as ‘start with a simple or familiar case’, ‘include uncommon cases’ and ‘keep unnecessary work to a minimum’ (Zodik & Zaslavsky, 2008). That is, replacing teachers’ current sequencing with interleaving is poorly warranted on the basis of studies where interleaving outperforms pure laboratory-style blocking.

Moreover, laboratory experiments appear to involve implicit, inductive learning which may not be the teachers’ intended outcome. Experimental participants acquire concepts from observation of sequences of exemplars and names alone and are not usually tested on explanations of category membership. School teaching – particularly in mathematics and science – is generally explicit and deductive. While rich, accurate concept images of mathematical categories are valuable, mathematical reasoning also relies on deductions from concept definitions (Tall & Vinner, 1981). Despite there also being no

classification advantage detected for exposition in their experiments, Rowlandson and Simpson (2023) argued for exposition: as well as being no worse for categorising, exposition students were better able to identify defining features of angle relationships.

Indeed, as well as involving implicit learning, the material in interleaving laboratory research is generally educationally irrelevant to participants: there is little value for psychology students in distinguishing Braque's skyscapes from Pessani's. The requirement for careful control leads researchers to use deliberately unfamiliar topics, so learning can be ascribed to training and not pre-existing knowledge. In contrast, school students are often working to extend and connect to pre-existing knowledge.

So, while knowledge brokers' argue interleaving is an effective classroom strategy by direct extrapolation from cognitive science research, this is not well founded. They conflate effects and the conditions which facilitate an interleaving/blocking effect in the laboratory are likely absent from classrooms.

### **Lab-ifying the World**

A second route to extrapolating is analogous to the process by which laboratory science becomes the basis for engineering: once experiments identify conditions under which a phenomenon is reliably produced, engineers can build appropriate technology to create those conditions and exploit the phenomenon. In contrast to the knowledge brokers' direct extrapolation above, in which experimental conditions are ignored, this second route involves their careful recreation.

Arguably Rowlandson and Simpson (2023) took this route, imposing lab-like conditions on school students to create the interleaving effect in the classroom. They were unsuccessful perhaps because they did not engineer just the right combination of conditions.

Current knowledge from interleaving experiments suggests an interleaved training phase will result in better classification than a blocked one for inductively learning to distinguish a small number of categories which form an exhaustive, non-overlapping partition of items in the field of interest, when there is a high degree of between-category similarity in a field for which judgement is an implicit, similarity based process, where there is little pre-existing knowledge of the field, where testing takes place soon after training and involves multiple choice responses, for adult participants with relatively low motivation to correctly classify and where there are no other intervening causes.

As the research field progresses, some of these features may come to be seen as unnecessary and others refined. Nonetheless, it may be possible to create the interleaving effect in the classroom if one tries hard enough to engineer a classroom situation with just these conditions.

But to what end?

Unlike researchers, teachers' aims are unlikely to include the creation of the interleaving effect; they are supporting learning of, say, angle relations in parallel lines, the structure of volcanoes or badminton serves. Just as Hacking argues that there is nowhere in nature that the Hall effect exists in its pure form, there is likely nowhere outside the laboratory that the interleaving effect exists in its pure form and no value to imposing so strongly on the classroom simply to generate a purer interleaving effect.

So, while technically possible to extrapolate via this route, engineering classrooms to create the interleaving effect probably has little practical value for teachers.

### **Mechanisms and Contexts**

The lack of an interleaving effect 'in nature' need not mean that the research has no value for education, just that knowledge brokers need a different route for extrapolating from the psychology laboratory to the classroom.

Much criticism of the 'what works' language of EBE is that the expensive field experiments cannot establish the kind of direct knowledge that one needs for policy (e.g. Joyce & Cartwright, 2020). Instead, it has been proposed that focus should shift to 'what works, for whom, in what circumstances'. One such approach is 'realistic evaluation' in which one tries to understand the mechanisms at work, the contexts in which they work (more or less successfully) and the outcomes (positive, neutral and negative, including side effects) that might result from those mechanisms acting in the given contexts (Pawson & Tilley, 1997).

While knowledge brokers have focussed on the treatment (interleaving), more success might come from focussing on the mechanisms at play across this research (e.g. discriminant contrast, commonality abstraction). In the case of interleaving, each of the mechanisms at play may form the basis for useful pedagogical interventions in the right circumstances, for a given outcome.

For example, if the intended outcome is merely improved classification, and if students confuse two categories because they are struggling to see what makes them different, a teacher might present examples sufficiently close together to enable discriminating features to be identified. That might involve interleaving, but may instead involve simultaneous presentation, asking students to list features, playing 'spot the difference' games etc. Alternatively, students struggling to identify the boundaries of a category because they cannot see what makes apparently different items cohere, might

benefit from seeing exemplars together so they can seek common features. That might be sequential pure blocking, but might also involve juxtaposition or asking students to explicitly list shared features or to generate their own examples and non-examples.

If the intended outcome is a particular approach to classification – such as being able to classify according to definitional rules – then perhaps these mechanisms are less appropriate.

### **Conclusions**

This paper has focussed on what can be taken from reports of experiments.

In discussing Harlow's experiments where orphaned monkeys spent more time with a soft cloth 'mother surrogate' which provided no food than a hard wire 'mother' which did provide food, Mook (1983) noted

Harlow did not conclude, "Wild monkeys in the jungle probably would choose terry-cloth over wire mothers, too, if offered the choice." First, it would be a moot conclusion, since that simply is not going to happen. Second, who cares whether they would or not? The generalization would be trivial even if true. What Harlow did conclude was that the hunger-reduction interpretation of mother love would not work. (Mook, 1983, p 381)

We should not take from Harlow's experiments that soft cloth mother surrogates "work". Unlike Hacking's Hall effect, which can be exploited to create a useful technology like a car speedometer, it is unclear how we can impose Harlow's laboratory conditions within a technology to usefully exploit the mother surrogate phenomenon. Instead, Harlow's work should be seen as contributing to a more abstract mechanism of attachment which might help influence practice less directly.



Brokered pedagogical knowledge, in the form of professional literature, attempts to take recommendations for practice from field and laboratory experiments through routes which may be as inappropriate as using Harlow's work to recommend soft cloth mother surrogates.

While the paper's focus has been on laboratory experiments, it does not contend that taking directly from field experiments provides much of a stronger foundation for educational policy or practice. Joyce and Cartwright (2020) argued that extrapolating from field experiments to the classroom often involves simple induction: it worked 'there', so it will work 'here'. There are many problems with this, not least that the difference in treatments in a field experiment involves a complex set of features which combine in unclear ways to produce an average positive outcome for the intervention group in the experimental context. A teacher might nonetheless judge the similarity of 'there' and 'here' to justify extrapolation. Field experiments in education may involve students of a similar age phase, with educationally relevant material, realistic tasks and outcome measures, medium to long term impact and (even if often poorly described) a control treatment similar to current school practice. The grounds for such extrapolation remain very weak, but knowledge brokers could focus on recommending the intervention treatment with strong caveats about context.

The focus here has been on extrapolating results from laboratory experiments, using the case of interleaving. Knowledge brokers appear to be taking the same approach to laboratory experiments as to field experiments: arguing by simple induction that interleaving works. In many cases they conflate spacing with interleaving effects. More importantly, the simple induction from 'it worked there' to 'it will work here' is even less plausible for laboratory experiments. The laboratory experiment better identifies the

causal factor at play, but one subject to the particular configuration of apparatus and measuring instruments in the study. For interleaving, participants are often psychology undergraduates, the material deliberately educationally irrelevant, the learning implicit, the impact measured short term and the comparison treatment is a pure blocked sequencing that few teachers or textbooks use as normal practice. The requirement for careful control means very few of the features of 'there' (the laboratory) will be present 'here' (the classroom), so direct extrapolation is even less justifiable than for field experiments.

There are two alternative extrapolation routes: intervening on the world to recreate exactly the right conditions to generate the effect, or using the theoretical knowledge of cognitive mechanisms. Through a controlled relaxation of control, Rowlandson and Simpson's (2023) experiments suggest that even a small deviation from the laboratory conditions can suppress the interleaving effect. Perhaps this occurred because of the nature of the mathematical categories and the judgement mechanisms for classical categorisation, or because of other subtle changes in experimental setup. Extending that sequence of extra-lab experiments might have eventually resulted in identifying the right configuration of features to create an interleaving effect. While that might have contributed to understanding those configurations, it would nonetheless have had little to say for practice.

Instead, knowledge brokers might focus on the theory about cognitive mechanisms generated by those experiments, identifying the contexts in which they work and how to exploit them.

Mixing together obviously different problems (as in Rohrer et al., 2014) is unlikely to improve category learning via the interleaving effect – though it may improve memory for

the procedures via the spacing effect. Instead, if students find it difficult to discern the extent of a single category, working with a variety of examples of that category together may help; if they struggle to tell some categories apart, working with examples from across those categories may help. Neither might be particularly effective if the outcome sought is the ability to identify categories from definitional features, which may be common in mathematics and other technical subjects.

Despite the evidence-based policy turn of the past three decades, we have yet to see education repeat the extraordinary success of the post-war turn towards evidence-based medicine. Knowledge brokers can be critical to future success, provided their work recognises that extrapolation in biology and in social sciences are distinct processes and that there are different pitfalls to extrapolating from field experiments and laboratory experiments. Brokering knowledge for practitioners may be more successful if it used laboratory experiments as sources of information about potential mechanisms and the contexts in which they work, than if it tries to impose laboratory conditions or continues to try to apply results from experiments directly to classroom practice.

### References

Agarwal, P. K., & Bain, P. M. (2019). *Powerful Teaching: Unleash the Science of Learning*.

John Wiley & Sons.

Baron, J. (2018). A brief history of evidence-based policy. *The Annals of the American*

*Academy of Political and Social Science*, 678(1), 40-50.

Barton, C. (2018). *How I Wish I'd Taught Maths: Lessons learned from research,*

*conversations with experts, and 12 years of mistakes*. John Catt Educational Ltd.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, 41(3), 392–402.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth.

Brown, P., C., Roediger, H., L., & McDaniel, M., A. (2014). *Make It Stick: The Science of Successful Learning*. Harvard University Press.

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, 145(11), 1029–1052.

Busch, B., & Watson, E. (2019). *The Science of Learning: 77 Studies That Every Teacher Needs to Know*. Routledge.

Cabinet Office. (2013). What works network. <https://www.gov.uk/guidance/what-works-network>.

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & cognition*, 41, 671-682.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

- Carvalho, P. F., & Goldstone, R. L. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. *Proceedings of the Annual Conference of the Cognitive Science Society*, 34, 186–191.
- Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(4), 1020-1031.
- Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General*, 127(4), 355-376.
- Dagenais, C., Lysenko, L., C. Abrami, P., M. Bernard, R., Ramde, J., & Janosz, M. (2012). Use of research-based information by school practitioners and determinants of use: A review of empirical research. *Evidence & Policy*, 8(3), 285-309.
- Davies, P. (1999). What is evidence-based education?. *British journal of educational studies*, 47(2), 108-121.
- Deans for Impact (2015). *The science of learning*. Austin, TX: Deans for Impact.
- Department of the Taoiseach. 2020. Programme for Government: Our Shared Future. <https://www.gov.ie/en/publication/7e05d-programme-for-government-our-shared-future/>.
- Edoald, T., & Nevill, C. (2021). Working out what works: The case of the Education Endowment Foundation in England. *ECNU Review of Education*, 4(1), 46-64.
- Eglington, L. G., & Kang, S. H. K. (2017). Interleaved Presentation Benefits Science Category Learning. *Journal of Applied Research in Memory and Cognition*, 6(4), 475–485.

Every Child Succeeds Act (ESSA) of 2015, P.L. 114-95, S.1177, 114th Cong. (2015).

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education, 9*(2), 642-684.

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Memory & Cognition, 47*(6), 1088–1101.

Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition, and education. *Trends in Neuroscience and Education, 4*(3), 49–59.

Gleeson, J., Harris, J., Cutler, B., Rosser, B., Walsh, L., Rickinson, M., Salisbury, M. & Cirkony, C. (2022). School educators' use of research: findings from two large-scale Australian studies. *Research Papers in Education, 1-25*.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Hargreaves, D . H . (1996) Teaching as a research-based profession: possibilities and prospects, Teacher Training Agency Annual Lecture 1996. London, Teacher Training Agency.

Higgins, S., Katsipataki, M., Villanueva Aguilera, A. B., Dobson, E., Gascoine, L., Rajab, T., Kalambouka, A., Reardon, J., Stafford, J., & Uwimpuhwe, G. (2022). The Teaching and Learning Toolkit: Communicating research evidence to inform decision-making for policy and practice in education. *Review of Education, 10*(1), e3327.

Hodgen, J., Barclay, N., Foster, C., Gilmore, C. Marks, R., & Sims, V. (2020) *Improving Mathematics in the Early Years and Key Stage 1: Guidance Report*. Education Endowment Foundation, London.

InnerDrive. (n.d.). What is interleaving, and why does it work?

<https://blog.innerdrive.co.uk/why-interleaving-works>

Innes, M. (2023). When policy intermediaries produce knowledge: A Bourdieusian analysis of the Education Endowment Foundation's influence in a multi-academy trust. *Journal of Education Policy*, 1-18 (online first).

Joyce, K. E., & Cartwright, N. (2020). Bridging the gap between research and practice: Predicting what will work locally. *American Educational Research Journal*, 57(3), 1045-1082.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative Contrast. *Applied Cognitive Psychology*, 26(1), 97–103.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction?” *Psychological Science*, 19(6), 585–592.

Lortie, Dan C. 2020. *Schoolteacher: A Sociological Study*. Chicago: University of Chicago Press.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human movement science*, 9(3-5), 241-289.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

- Metcalfe, J., & Xu, J. (2016). People mind wander more during massed than spaced inductive Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(6), 978-984.
- Nagatsu, M., & Favereau, J. (2020). Two strands of field experiments in economics: A historical-methodological analysis. *Philosophy of the Social Sciences*, 50(1), 45-77.
- No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).
- Nunes, T., Barros, R., Evangelou, M., Strand, S, Mathers S. & Sanders-Ellis, D. (2018) *1stClass@Number: Evaluation report and executive summary*, Education Endowment Foundation, London.
- Pawson, R., & Tilley, N. (1997). *Realistic Evaluation*. Sage.
- Productivity Commission (2016). *National Education Evidence Base: Productivity Commission Inquiry Report: Overview and Recommendations*. Productivity Commission. Canberra: Australian Government.
- Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, 35(6), 481–498.
- Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, 21(5), 1323–1330.
- Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, 4(3), 328-350.



Rowlandson, P. & Simpson, A. (2023) Interleaving in mathematical category learning.

PsyArXiv. <https://doi.org/10.31234/osf.io/gz5r7>

Rycroft-Smith, L. (2022). Knowledge brokering to bridge the research-practice gap in education: Where are we now?. *Review of Education*, 10(1), e3341.

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, 109(1), 84–98.

Shadish WR, Cook TD, & Campbell DT (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15-21.

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, 55(1), 21-31.

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, 12(2), 151-169.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, 24(6), 837–848.

Torphy, K., Liu, Y., Hu, S., & Chen, Z. (2020). Sources of professional support: Patterns of teachers' curation of instructional resources in social media. *American Journal of Education*, 127(1), 13-47.

- Vlach, H. A., & Sandhofer, C. M. (2013). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development*, 83(4), 1137–1144.
- Weinstein, Y., Sumeracki, M., & Caviglioli, O. (2018). *Understanding How We Learn: A Visual Guide*. Routledge.
- Zodik, I., & Zaslavsky, O. (2008). Characteristics of teachers' choice of examples in and for the mathematics classroom. *Educational Studies in Mathematics*, 69(2), 165-182.
- Zulkipli, N. (2013). Effect of interleaving exemplars presented as auditory text on long-term retention in inductive learning. *Procedia - Social and Behavioral Sciences*, 97, 238–245.
- Zulkipli, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, 41(1), 16–27.



**Citation on deposit:** Rowlandson, P., & Simpson, A. (in press). Brokering Knowledge from Laboratory Experiments in Evidence-Based Education: The Case of Interleaving. *British Educational Research Journal*

**For final citation and metadata, visit Durham Research Online URL:**

<https://durham-repository.worktribe.com/output/2442988>

**Copyright statement:** This accepted manuscript is licensed under the Creative Commons Attribution 4.0 licence.

<https://creativecommons.org/licenses/by/4.0/>